ENRICHING AND ENHANCING MOVING IMAGES WITH LINKED OPEN DATA: OPPORTUNITIES AND IMPLEMENTATION CHALLENGES

Karen F. Gracy

School of Library and Information Science

Kent State University

JUMPING ON THE BANDWAGON! THE LINKED OPEN DATA MOVEMENT

- LOD is all about connecting related data that is currently isolated ("siloed") in various cultural heritage repository information systems to a larger universe of related data (not just data in LAM repositories)

- Implementation of LOD will enable LAMs to …

  - ✓ Enhance their already existing cataloging records, finding aids and other services;
  - ✓ Help users obtain new information and knowledge relevant to materials in their institution; and,
  - ✓ Make their collections available to users of external databases.

MAJOR IMPEDIMENTS TO IMPLEMENTING LINKED OPEN DATA FOR MOVING IMAGES

- LAMs often create datasets and systems that keep records of objects and collections, but …
  - the various pieces of information about creators, places, and events found in each record are not easily connected to relevant information in other sources.
  - To make those connections, alignment of the various vocabularies and schemas must be accomplished.
- Information systems use metadata schemas and vocabularies that may not be easily aligned to one another, for the following reasons:
  - Mismatch in granularity between vocabularies
  - Hidden information in unstructured fields
- Much important information relating to moving images is proprietary and not likely to be made accessible by the companies for whom metadata aggregation and sales are their business.
  - Open data sources such as DBPedia and Freebase should be explored as alternatives, with the caveat that not all information may be as complete or "authoritative" as we might like.

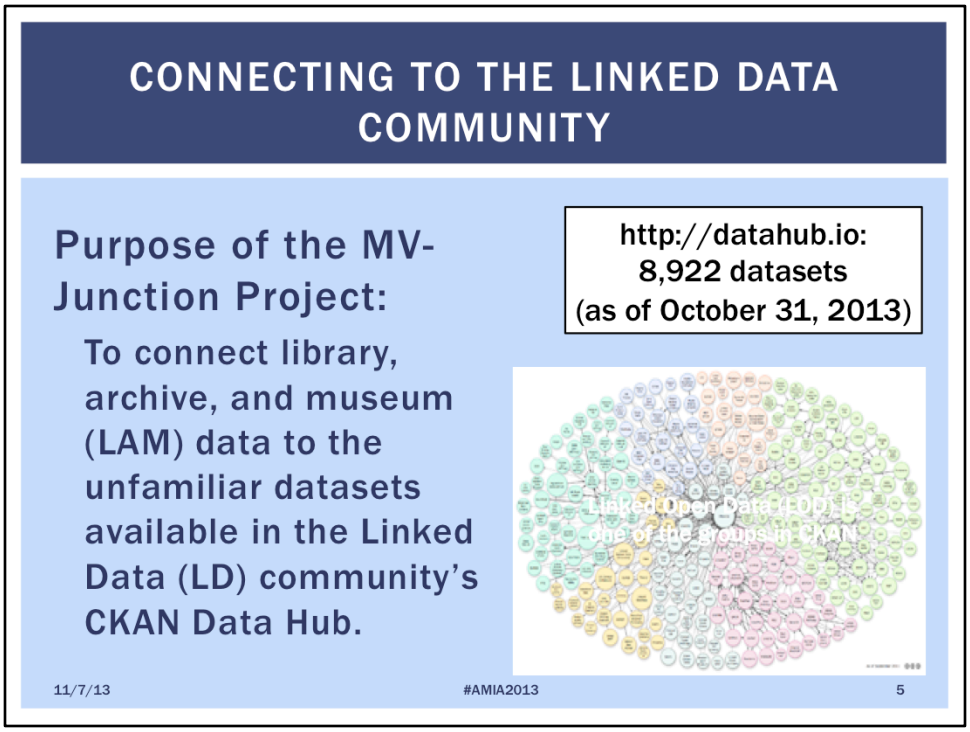
RECENT RESEARCH ON LINKED OPEN DATA FOR LAMS AT KENT STATE UNIVERSITY

The Metadata Vocabulary Junction Project:

Connecting Library Data to the Unfamiliar Data and Metadata Resources in the LOD Universe

PI: Marcia Lei Zeng, Ph.D.
Co-PI: Karen Gracy, Ph.D.
Research Assistants:
Laurence Skirvin, MLIS; Sammy Davidson, MLIS; Shadi Shakeri, MLIS; Riley Stoermer, MLIS

http://lod-lam.slis.kent.edu/


CONNECTING TO THE LINKED DATA COMMUNITY
Purpose of the MV-Junction Project:
To connect library, archive, and museum (LAM) data to the unfamiliar datasets available in the Linked Data (LD) community's CKAN Data Hub.
http://datahub.io:
8,922 datasets
(as of October 31, 2013)
11/7/13
#AMIA2013
5

This presentation will report on a research study being conducted on how archives, libraries, and other cultural institutions can benefit from the data and metadata resources that have been created as a result of the Linked Open Data (LOD) movement. The LOD movement aims to connect related data formerly isolated in repositories (often called "silos") and not previously linked. Archival institutions generate valuable data through tools such as finding aids, cataloging records, and databases, but this information is often sequestered inside archival information systems that cannot be easily linked to other data sources. Linked data technologies and LOD resources can enable archives to enhance their already existing finding aids and other services, allow archival users to obtain new information and knowledge relevant to materials in the archive, and make available archival collections to the users of external databases.

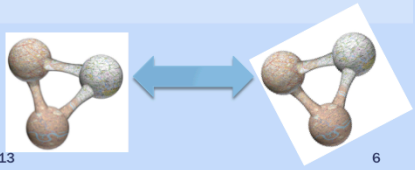
ALIGNING LAM DATA AND LOD DATA CONSTRUCTS

LAM data:

- Library bibliographic records MARC records
  - schema.org markup
- Digital collection metadata
  - Dublin Core-based
  - Other locally defined
- Archival descriptions
  - EAD, MARC, PBCore, others
- Museum and Visual Resources object descriptions
  - VRA Core, others

LOD Datasets:

- Ontologies
- Metadata Application Profiles
- Structured data in XML/RDF samples
- Documentation



FOCUS AREAS

- Authority information
- Music and music-related information
- Archives and archival finding aids
- Museum object documentation
- Theses and dissertations
- Moving images and related information (currently in progress)

EXAMPLE 1: MUSIC-FOCUSED STUDY

DATA COLLECTION AND ANALYSIS (1)

LOD Datasets:

- Identify music-related datasets
- Analyze the source of the data structures
  - Ontologies used
    (e.g., Music Ontology, Similarity Ontology, Event Ontology, Programmes Ontology, etc.)
  - Metadata schemas and application profiles
    (e.g., BBC Music Schema and derivatives)
  - Structured data in XML/RDF samples
  - Documentation
- Crosswalk their properties, indicating the matching level
  - (e.g., broadMatch, equivelent, narrowMatch, etc.)
- Identify major classes and properties useful to library data
  - (e.g., *mo:MusicArtist, rev:Review, mo:Performance, mo:performer, cc:license , dc:title, foaf:primaryTopic*, etc.)

EXAMPLE 1: MUSIC-FOCUSED STUDY

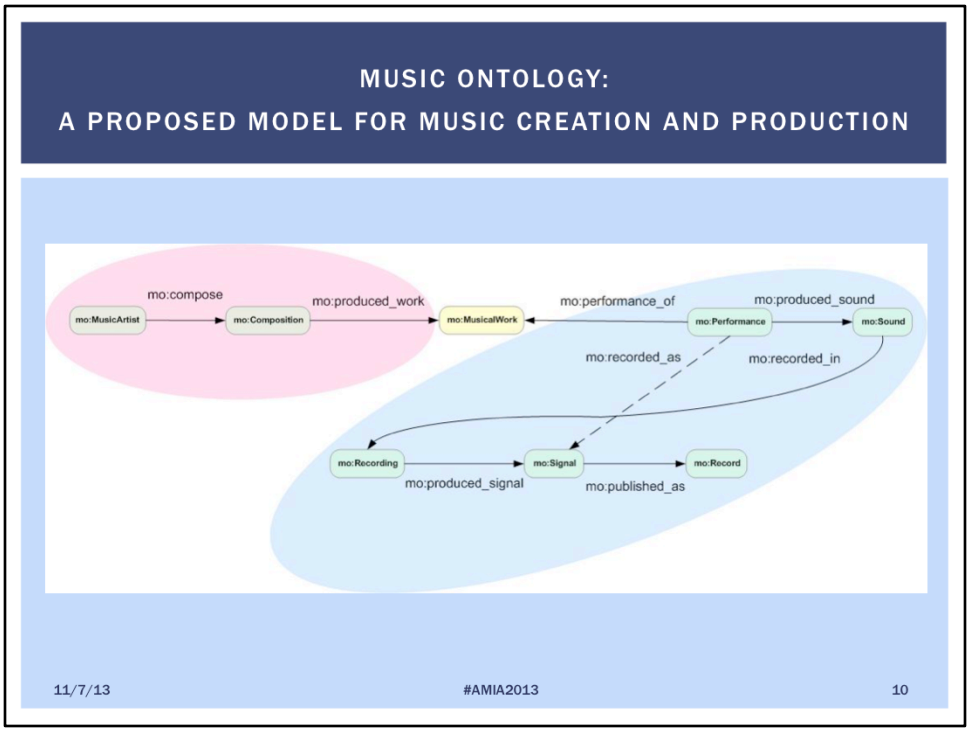
DATA COLLECTION AND ANALYSIS (2)

LAM data from:

1. Library MARC sample records
   - Classical/Instrumental
   - Jazz
   - Musical and Opera
   - Pop/Rock
   - Soundtrack
2. Their schema.org markup in the WorldCat
3. Digital collections' metadata structure
   20 digital collections from national libraries, academic libraries, achieves
   - More types (e.g., sheet music)

Break down the 'records' and find:

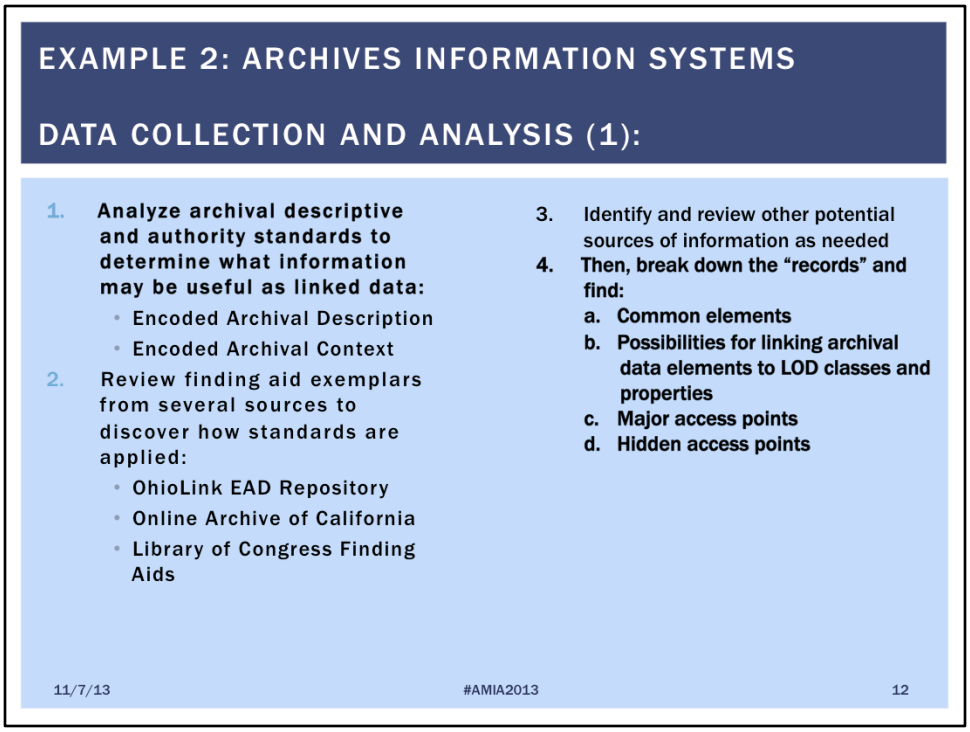
- Common elements
- Possibility of linking to LOD properties
  - Major access points
  - Hidden access points



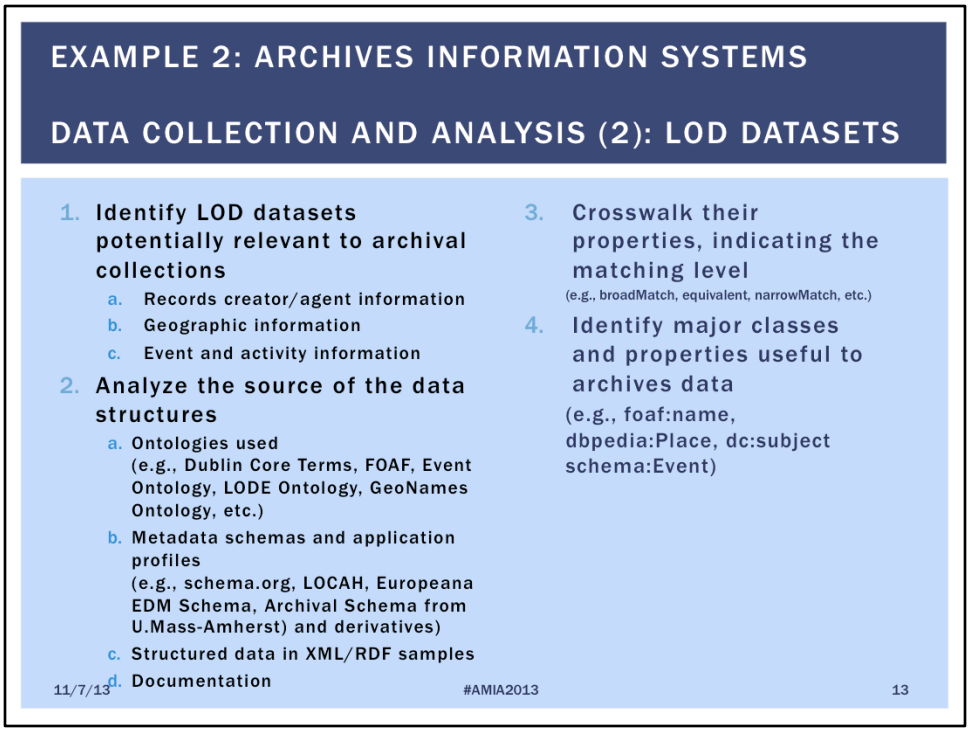
We looked closely at this model before and while doing our vocabulary alignment work. Notice that distinctions are made between various events: Composition, Performance, Sound, Recording, Signal, and Record are all classes of events that related to the Musical Work itself.

ALIGNMENT EXAMPLE: MARC TO MUSIC ONTOLOGY

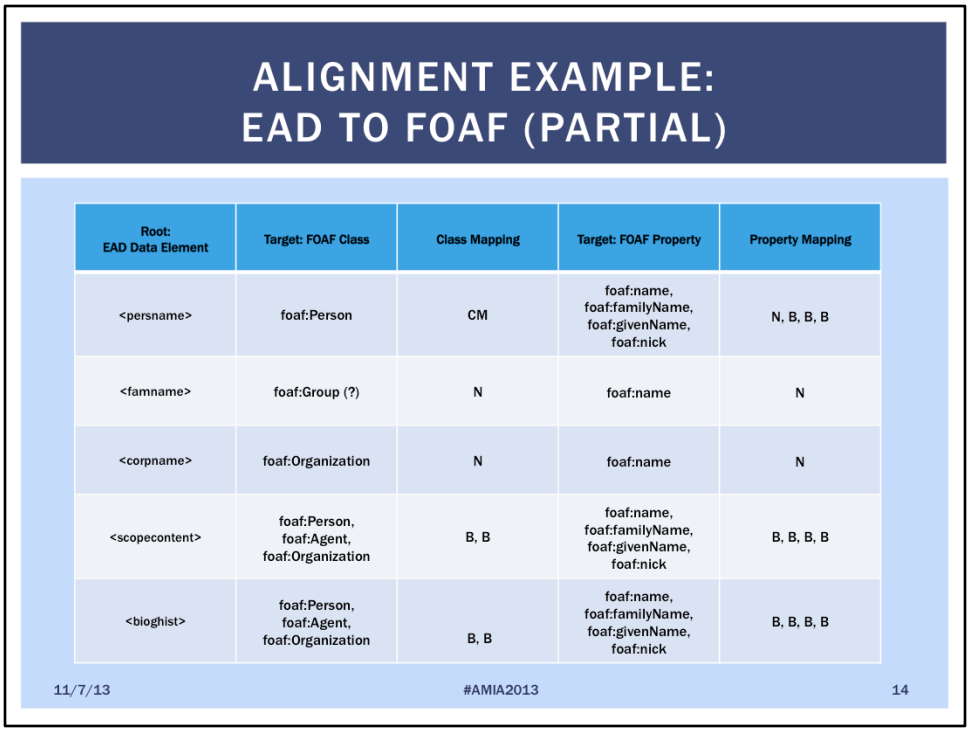
Root: MARC Field/Subfield	Target: MO Class	Class Mapping	Target: MO Property	Property Mapping
245 / $a (Title Proper)	mo:Record	B	dct:title	CM
700 / $t (Title of a work)	mo:MusicalWork	B	dct:title	CM
505 / $a (Content Note)	mo:MusicalWork, mo:Track	B	mo:track, dc:description	B, N
100 / $a (Personal Name)	mo:MusicArtist	B	mo:performer, mo:singer, mo:composer mo:conductor	B, B, B, B
650 / $a (Subject Headings)	mo:Genre	R	mo:genre, dct:type	R, R
655 / $a (Genre/Form Terms)	mo:Genre	R	mo:genre, dct:type	CM, N


EXAMPLE 2: ARCHIVES INFORMATION SYSTEMS
DATA COLLECTION AND ANALYSIS (1):
1. Analyze archival descriptive and authority standards to determine what information may be useful as linked data:
• Encoded Archival Description
• Encoded Archival Context
2. Review finding aid exemplars from several sources to discover how standards are applied:
• OhioLink EAD Repository
• Online Archive of California
• Library of Congress Finding Aids
3. Identify and review other potential sources of information as needed
4. Then, break down the "records" and find:
a. Common elements
b. Possibilities for linking archival data elements to LOD classes and properties
c. Major access points
d. Hidden access points
11/7/13
#AMIA2013
12

Our research process includes two major steps: for the first part, we have been examining archival descriptive and authority standards to see what information may be useful as linked data. We are also looking at finding aid and authority exemplars, which we pulled from major sources of archival finding aids such as the OhioLink EAD repository. After reviewing standards and sample finding aids and authority records, we identify common elements among them and consider the possibilities for linking archival information to linked open data properties—we are looking at both major access points and potential hidden access points not currently able to be linked easily because of how they are currently encoded.


EXAMPLE 2: ARCHIVES INFORMATION SYSTEMS
DATA COLLECTION AND ANALYSIS (2): LOD DATASETS
1. Identify LOD datasets potentially relevant to archival collections
a. Records creator/agent information
b. Geographic information
c. Event and activity information
2. Analyze the source of the data structures
a. Ontologies used (e.g., Dublin Core Terms, FOAF, Event Ontology, LODE Ontology, GeoNames Ontology, etc.)
b. Metadata schemas and application profiles (e.g., schema.org, LOCAH, Europeana EDM Schema, Archival Schema from U.Mass-Amherst) and derivatives)
c. Structured data in XML/RDF samples
d. Documentation
3. Crosswalk their properties, indicating the matching level
(e.g., broadMatch, equivalent, narrowMatch, etc.)
4. Identify major classes and properties useful to archives data
(e.g., foaf:name, dbpedia:Place, dc:subject schema:Event)
11/7/13
#AMIA2013
13

In the second part of the research, we have identified datasets that have potential to be relevant to users of archival collections. These datasets are analyzed to discover what ontologies, metadata schemas, and/or application profiles that are used. We also look at sample records and any documentation that we can find, using information found on the Data Hub. Then, we attempt to crosswalk across datasets, to see how well a particular type of information will match. We use categories common to ontology alignent such as equivalent, close match, broad match, narrow match, and related. Through this process, we identify the major classes and properties useful to archives data.

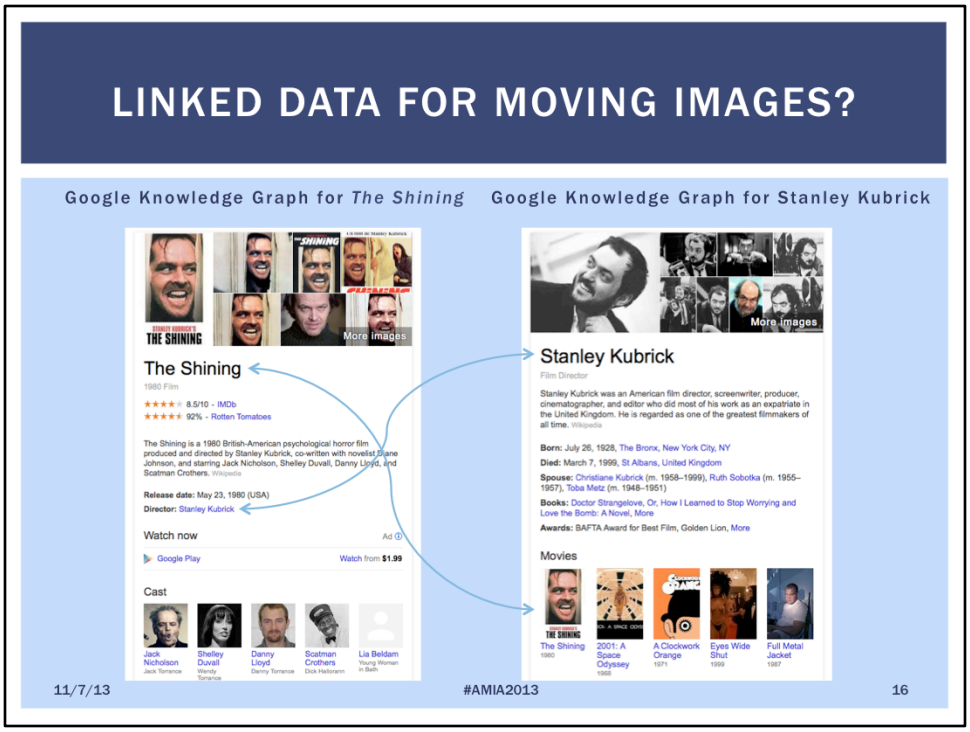


This is one example of a mapping table that is generated as a result of the analysis we're doing with the standards, ontologies, and datasets. In this mapping of EAD to FOAF, which focuses on personal, family, and corporate body names, we can make a couple of general statements about how well these two data models relate. First, we can say that for the EAD data elements <persname>, <famname>, and <corpname>, at the class level they are narrower than the closest FOAF class equivalent, except in the case of Person. Matching to the property level, however, FOAF tends to be narrower in scope than the closest equivalent in EAD. Matching names found in EAD records to the property foaf:name would provide a basic level of interoperability, but would not differentiate among first, last, and nicknames. Thus, EAD and FOAF are not completely equal in the granularity of their data models.

Another observation that we have made in EAD is that there are several places in the EAD record where names can be found, but which are not and cannot be tagged as such, such as in the ScopeContent and BiogHist tags. The potential to link additional names to external datasets is therefore untapped at this time, but revisers of the standard may wish to consider adding such functionality to future versions of EAD.

ALIGNMENT EXAMPLE: EAD TO SCHEMA.ORG (PARTIAL)

Root: EAD Data Element	Target: schema.org Class or *property*	Class/Property Mapping
<persname>	schema:Person	CM
<famname>	schema:Agent	B
<corpname>	schema:Organization	N
<scopecontent>	*schema:description*	B
<bioghist>	*schema:description*	CM

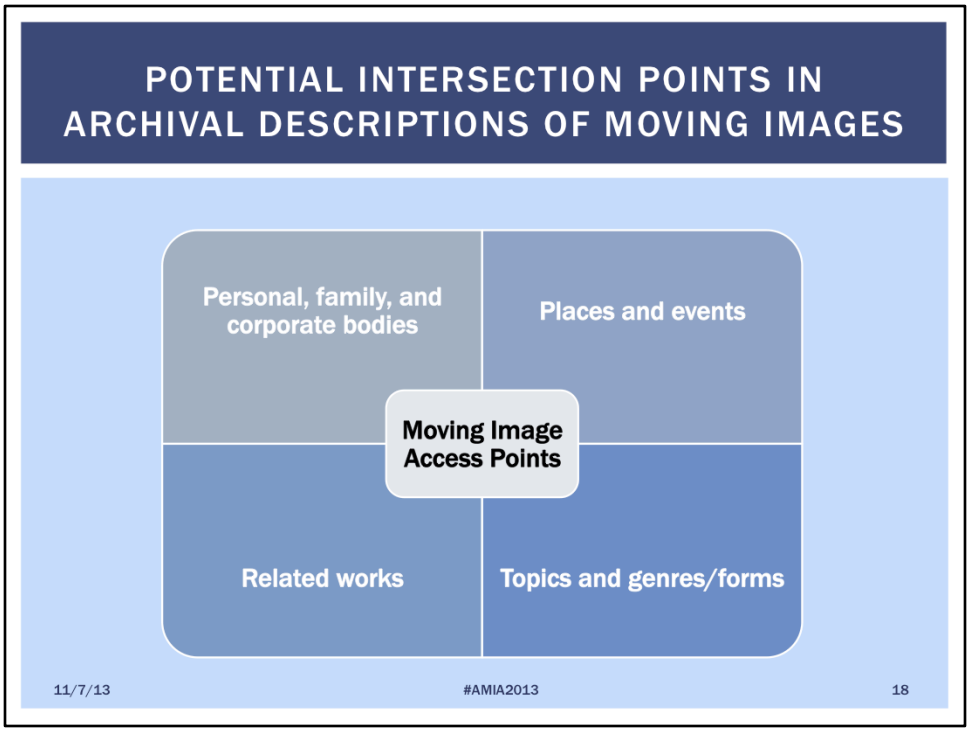

LINKED DATA FOR MOVING IMAGES?
Google Knowledge Graph for The Shining
Google Knowledge Graph for Stanley Kubrick
More images
The Shining
1980 Film
8.5/10 - IMDb
92% - Rotten Tomatoes
The Shining is a 1980 British-American psychological horror film produced and directed by Stanley Kubrick, co-written with novelist Diane Johnson, and starring Jack Nicholson, Shelley Duvall, Danny Lloyd, and Scatman Crothers. Wikipedia
Release date: May 23, 1980 (USA)
Director: Stanley Kubrick
Watch now
Ad
Google Play
Watch from $1.99
Cast
Jack Nicholson
Jack Torrance
Shelley Duvall
Wendy Torrance
Danny Lloyd
Danny Torrance
Scatman Crothers
Dick Hallorann
Lia Beldam
Young Woman in Bath
More images
Stanley Kubrick
Film Director
Stanley Kubrick was an American film director, screenwriter, producer, cinematographer, and editor who did most of his work as an expatriate in the United Kingdom. He is regarded as one of the greatest filmmakers of all time. Wikipedia
Born: July 26, 1928, The Bronx, New York City, NY
Died: March 7, 1999, St Albans, United Kingdom
Spouse: Christiane Kubrick (m. 1958–1999), Ruth Sobotka (m. 1955–1957), Toba Metz (m. 1948–1951)
Books: Doctor Strangelove, Or, How I Learned to Stop Worrying and Love the Bomb: A Novel, More
Awards: BAFTA Award for Best Film, Golden Lion, More
Movies
The Shining
1980
2001: A Space Odyssey
1968
A Clockwork Orange
1971
Eyes Wide Shut
1999
Full Metal Jacket
1987
11/7/13
#AMIA2013
16

Triples might be (e.g.):

The film The Shining was directed by Stanley Kubrick.
Stanley Kubrick directed The Shining.
The actor Jack Nicholson appeared in The Shining.
The Shining was released on May 23, 1980.

EXAMPLE 3 (ONGOING): ALIGNING VOCABULARIES FOR MOVING IMAGE DESCRIPTIONS

- Vocabularies/metadata element sets
  - MARC, PBCore, EAD
  - Ontology for Media Resources
  - Broadcast standards: MPEG-7, SMPTE Data Dictionary
- LD Datasets in the DataHub with moving image-related data:
  - DBPedia
  - Freebase
  - IMDB (not open data, but some use is possible)
    - Bulk dump files only (would require some data wrangling to clean up for use)
  - Linked Movie Database
    - 6,148,121 triples, mixture of material from DBpedia, Freebase and Geonames
- Closed datasets that might be useful if openly available as linked data:
  - Metadata created for cable/satellite electronic program guides (IPTV)
    - Currently owned by 3rd party metadata aggregators such as Tribune TV Data, FYI Television, and Rovi in the U.S. and Canada, TV-Anytime in Europe and Japan.

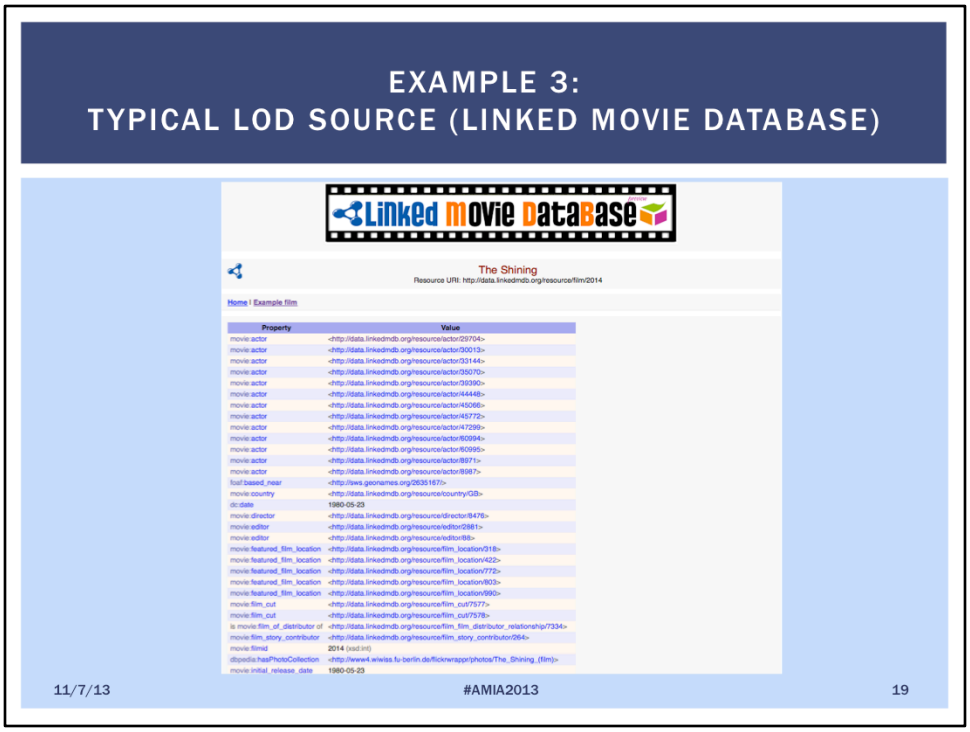


At this stage of our research, we are focusing on data elements, classes, and properties in four major categories:

1. Personal, family, and corporate bodies
2. Places and events
3. Related works
4. Topics and genres/forms

We see the potential for linking from archival descriptions in these types of things or entities to numerous data sets. Some examples of datasets that could prove helpful for archives users include:

1. As we have already seen from Daniel, VIAF is an invaluable source of linked open data for personal, family, and corporate body names.
2. DBPedia and Freebase are very large sources of contextual information about moving images, particularly commercially produced moving images.
3. Linked MDB is a film-focused LOD dataset that draws from DBPedia and Freebase, among others.
4. GeoNames: The GeoNames geographical database is available for download free of charge under a creative commons attribution license. It contains over 10 million geographical names and consists of over 8 million unique features



Each line points to the URI for a nugget of information (aka a triple). For example, the first value is for the actor Jack Nicholson, who starred in the movie The Shining.

PARTIAL ALIGNMENT SAMPLE: MARC TO LINKEDMDB

Root: MARC Field	Target: LMDB Property	Property Mapping
245 (Title Proper)	movie:film	CM
508 (Creation/Production Credits Note)	movie:director	N
511 (Participant or Performer Note)	movie:actor	N
655 (Genre/Form Terms)	movie:film_genre	CM
257 (Country of Producing Entity)	movie:country	CM

EXAMPLE 3: POTENTIAL INTERCHANGE VOCABULARY (ONTOLOGY FOR MEDIA RESOURCES)

- "The intent of this vocabulary is to bridge the different descriptions of media resources, and provide a core set of descriptive properties." Ontology for Media Resources 1.0, W3C Recommendation, 09 February 2012 http://www.w3.org/TR/2012/REC-mediaont-10-20120209/
- Documentation for onotology has pre-defined mappings for various vocabularies used in media description, including:
  - Dublin Core
  - EBUCore
  - MPEG-7
  - OGG
  - QuickTime
  - Various metadata aggregators for cable/satellite guides

PARTIAL ALIGNMENT SAMPLE: PBCORE TO ONTOLOGY FOR MEDIA RESOURCES

Root: PBCore Class/Property	Target: OMR Property	Property Mapping
pbcoreTitle	ma:title	CM
pbcoreSubject	ma:film_subject	CM
pbcoreCreator/creator	ma:creator	CM
pbcoreCoverage	ma:location	N
pbcoreDescription	ma:description	CM

Ontology for Media Resources may serve as an intermediary for alignment of vocabularies for moving images in the same way that the Music Ontology serves that role for musical works.

CHALLENGES FOR IMPLEMENTING MOVING IMAGES LOD (1)

- Similar to music and archives description, "hidden information" and hidden potential access points abound in moving image records, such as:
  - Title information
    - May be found in MARC 130, 245, 246, 505, 730, and 740
    - May be found in Dublin Core and PBCore description fields
  - Cast and crew information
    - Director name can be found in MARC 130, 245, 505, 508, 700)
    - DC and PBCore description fields may contain some contributors not given separate treatment in the Contributor or Creator fields.
  - Important technical information pertaining to a particular edition of a work may be found in unusual places, such as:
    - Aspect ratio (found in various MARC fields: 250, 500, 505, 538)
  - Multi-work records present a particularly difficult problem (how do you easily dig out each information nugget when there are multiple potential triples in a single field?).

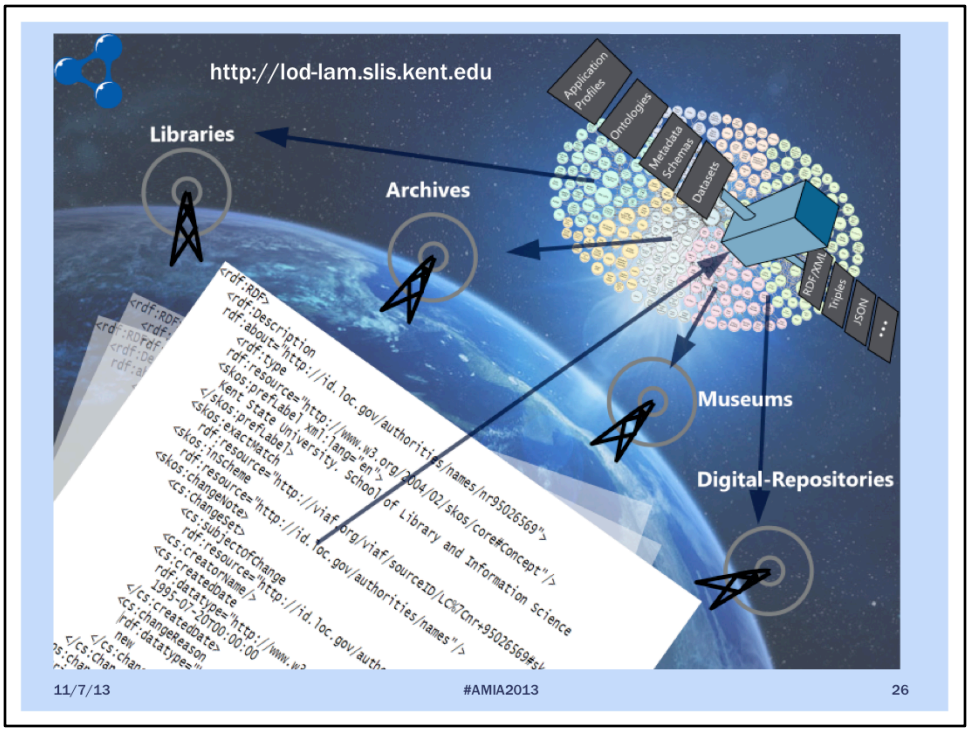
CHALLENGES FOR IMPLEMENTING MOVING IMAGES LOD (2)

- Lack of granularity in commonly used metadata element sets for information discovery (e.g., Dublin Core, PBCore) makes it difficult to align these to more granular sets of elements found in production-based vocabularies such as SMPTE Data Dictionary and MPEG-7.
  - Lack of semantic specificity in defining our data hinders our ability to share our information and connect to other sources of information.
- Do we need to develop a model for moving image production and archiving that identifies key events, agents, and processes?
  - Having such a model may help us in aligning current vocabularies and developing new ones that more closely address our needs (we're not really producers, so those vocabs developed for controlling production elements do not consider the full lifecycle of archiving).

THANK YOU!

- Questions after the presentation?
  - Please contact:

Karen F. Gracy
School of Library and Information Science
Kent State University
kgracy@kent.edu
http://lod-lam.slis.kent.edu


http://lod-lam.slis.kent.edu
Libraries
Archives
Museums
Digital-Repositories
Application Profiles
Ontologies
Metadata Schemas
Datasets
RDF/XML
Triples
JSON
...
11/7/13
#AMIA2013
26